Environmental Genome Shotgun Sequencing of the Sargasso Sea

J. Venter, K. Remington, J. Heidelberg, A. Halpern, D. Rusch, J. Eisen, D. Wu, I. Paulsen, K. Nelseon, W. Nelson, D. Fouts, S. Levy, A. Knap, M. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H.Baden-Tillson, C. Pfannkoch, Y. Rogers, H. Smith

Science 2004

Presented by Kevin Liu, UTCS

February 2009

New era of genomics

- Older sequencing technology: Sanger
  - Targeted read specified by primer
  - 300-1000b per read
- Newer sequencing technology: Whole-genome shotgun sequencing
  - Random reads across all genome
  - Automated fragment assembly of entire genome
  - May or may not use reference genome for assembly

Shotgun Sequencing

- Actually not that new: in use in 1979 on small genomes (Staden 1979)
- Human Genome Project
- Fueling growth of sequence database, which are growing at a pace faster than Moore's law (Goldman 2008)
  - New biochemistry, e.g. Pyrosequencing, vectors
  - Massively parallelizable because random reads are independent, and sequence assembly from random reads is a divide and conquer technique!

Shotgun Sequencing

- Technology advance much like transistor advances of the 70s, 80s, 90s, 2000s
  - 1-2 orders of magnitude faster
  - 1-2 orders of magnitude cheaper
  - 1-2 orders of magnitude shorter reads
  - (no citation – need to verify)
- Motivates two biological problems
  - Cancer/disease discovery via genome comparison
  - New microbiological species discovery

New Grand Challenges in Medicine and Genome Comparison

- Mapped reads of a new individual genomes against reference genome to detect evolutionary events
- 1000 Genomes Project
- Human Microbiome Project
- Cancer Genome Atlas Project
- Reinhoff's daughter (Wired, 2009)

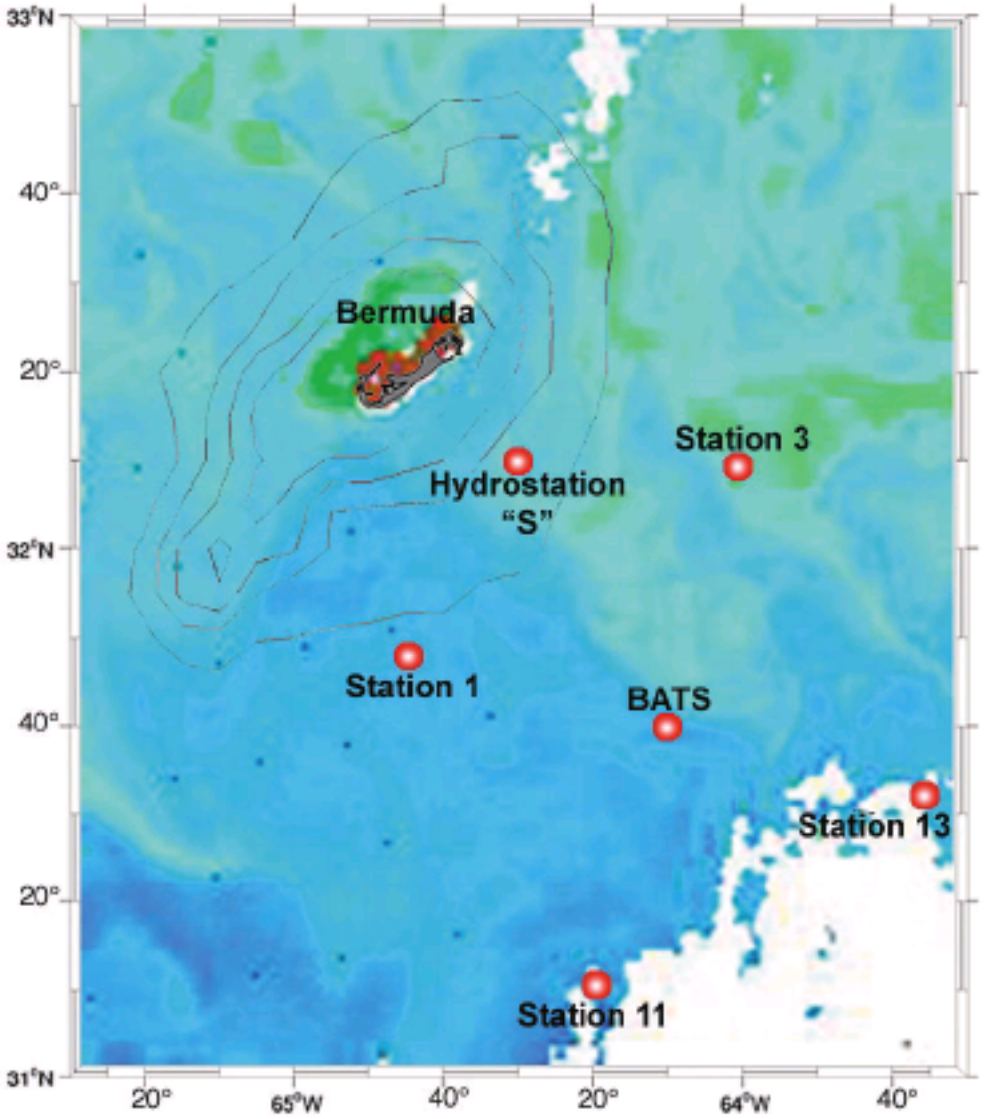
Metagenomics for Microbiology

- Venter 2004 paper is nice example
- Multiple areas of computational biology all rolled into one problem
  - Phylogenetics
  - Whole genome random short read sequencing problems
  - Biogeography
  - Population genetics
  - Ecology

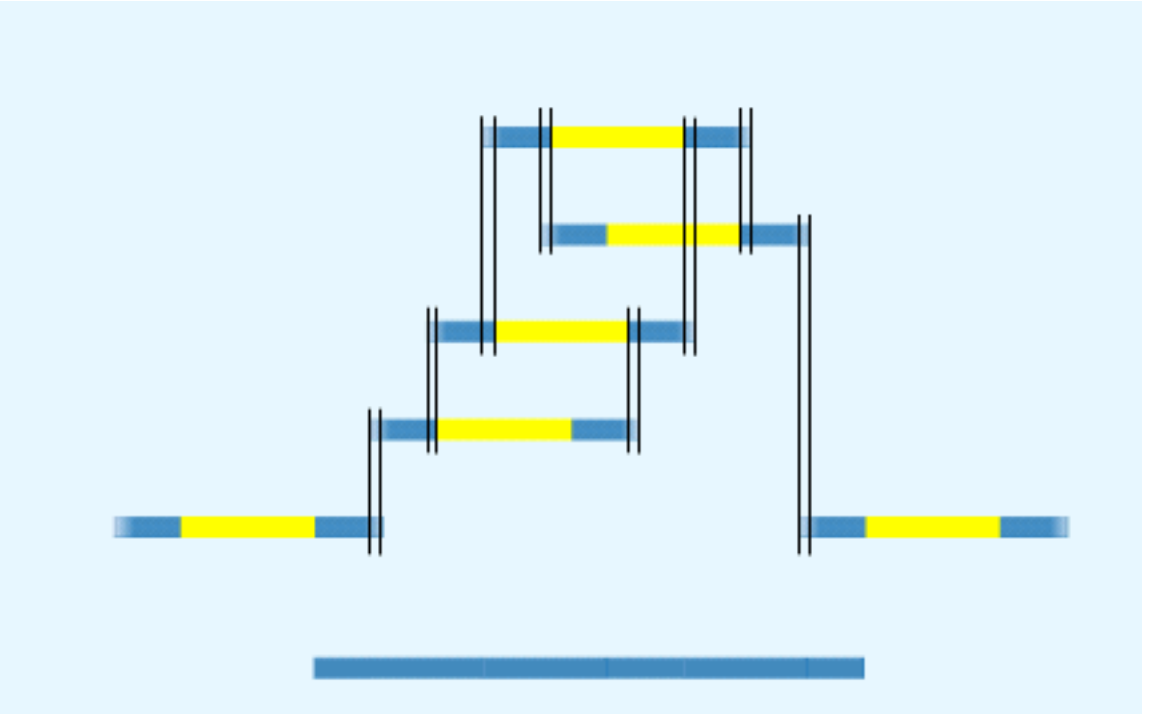
Metagenomics for Microbiology

- What if there is no known mapped genome for a species?
- What if it's difficult to find or detect individuals from a new species?
- What if individuals from a new species can't be cultured in a lab?
- Environmental sample shotgun sequencing

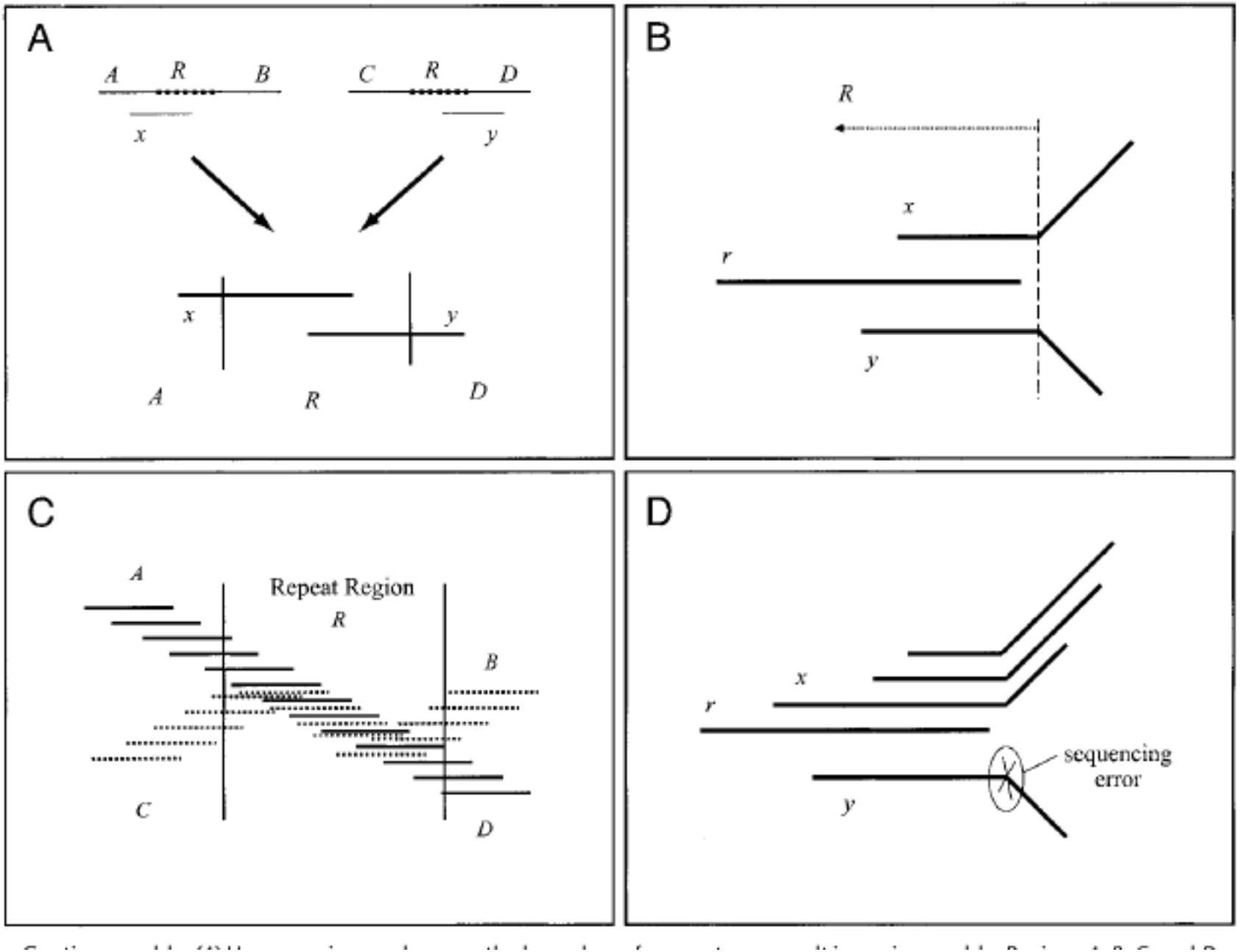
“Gallon” of Seawater from Bermuda



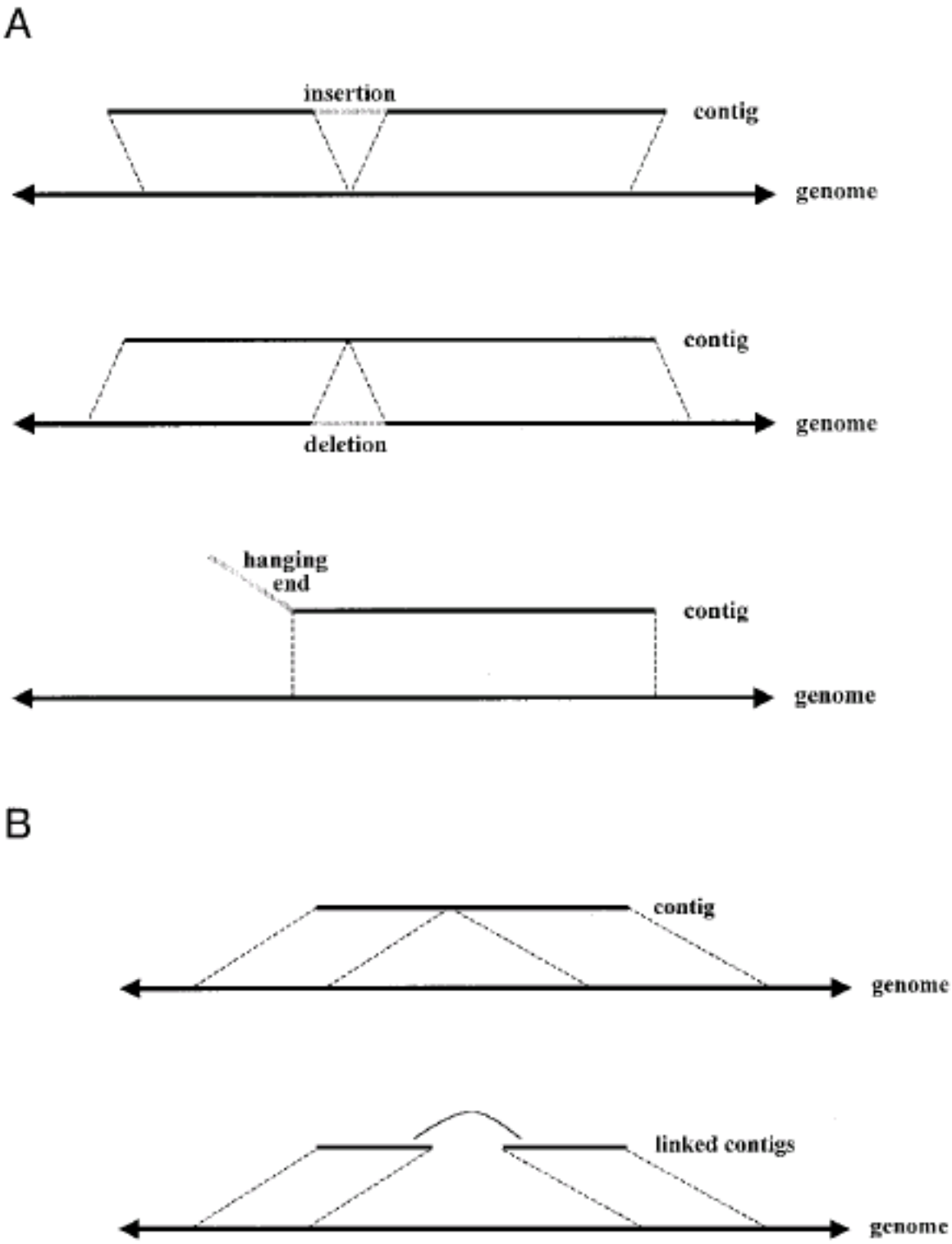
Short Random Read Fragment Assembly



Assembly Problems

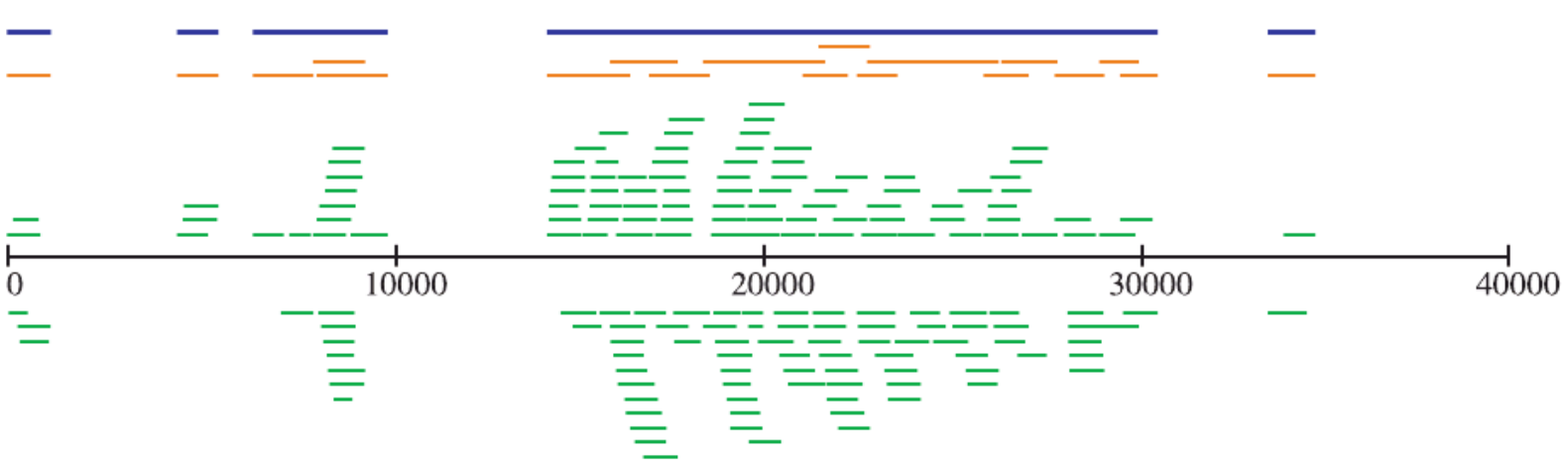


Mapped Assembly Problems



Venter 2004 Assembly Example

A



B

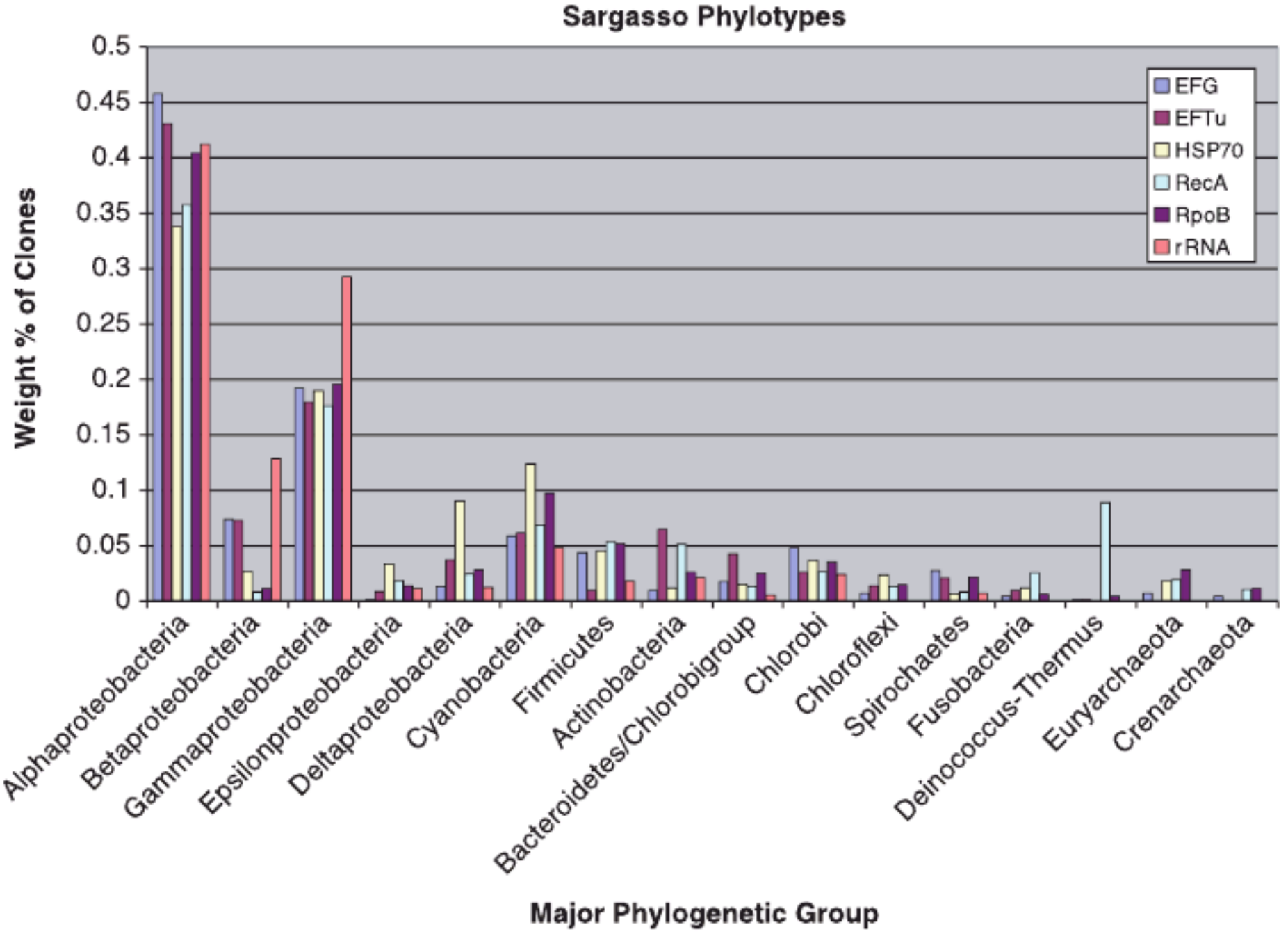
```
                                  <<<  Contig 856008, gapped length: 18972  >>>
2701 gapped
         |         |         |         |         |         |         |         |         |         |
2690 ungapped
u         u         u         u         u         u         u         u         u         u
CGGATAGTGTCTTATTAGATGAACTTAGTCCTGAAATAATCAAAAAAAACGTTCAAGAATTAAGAGAGTTATCAATTAATAGTCTACAAAAAGAAGTCAA  cns   (uid,iid) type
lllllllllllllllllllllllllllllllllllllllllllllllllllllllllllllllllllll@lllllllllllllKMll;lllllllllllll  qlt
```

```
.....................................................................................  >   (58629,58331)
                                                                                       >   (58629,58331)
.....................................................................................  >   (845377,839631)
                                                                                       >   (845377,839631)
.....................................c...........g...................t..............ct..a.............  >   (1408534,1338218)
                                     _           _                   X                a_  _              >   (1408534,1338218)
.....................................................................................  <   (1683713,1613397)
                                                                                       <   (1683713,1613397)
.a.............g...........c.....................................t........g.....ct..a..........a..  >   (831801,826055)
 Y              _           \                                    V        _     __  T          _     >   (831801,826055)
.a.............g...........c.....................................t........g.....ct..a..........a..  <   (1708001,1637685)
 V              J           X                                    _        S     BB  \          \     <   (1708001,1637685)
```

Venter 2004 Mapped and De Novo Assembly

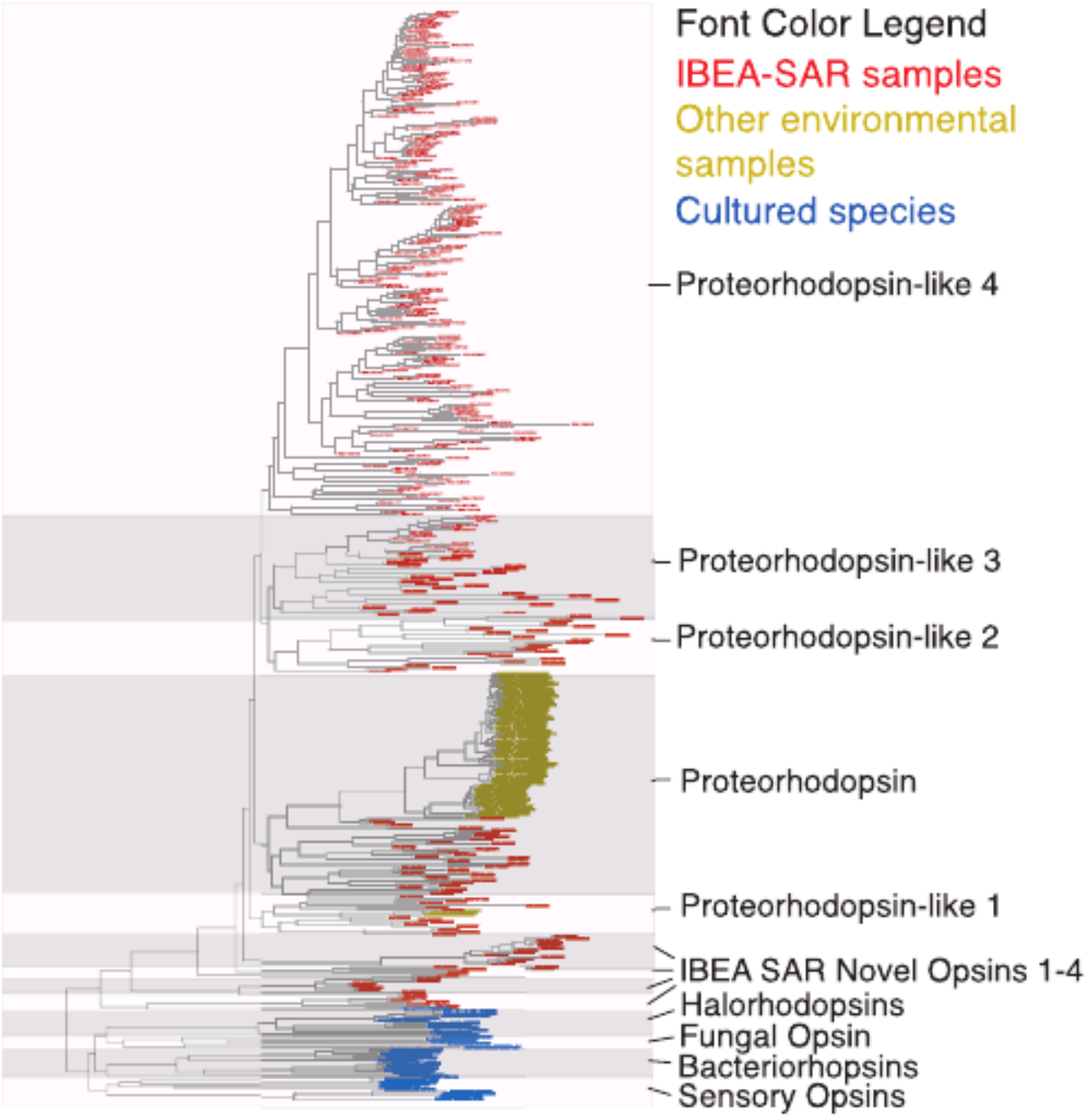
- Start by mapping against known genomes, markers
  - BLAST to GenBank
- New species discovery
  - “Sequence similarity is not necessarily an accurate predictor of functional conservation and sequence divergence does not universally correlate with the biological notion of 'species', defining species (also known as phylotypes) by sequence similarity within the rRNA genes is the accepted standard in studies of uncultured microbes

Venter 2004 Phylogenetic Groups by Marker



Venter 2004 Single Marker Tree

- ClustalW alignment, too many sequences for manual adjustment!
- Protdist and then Neighbor-joining tree



Venter 2004 True Species Counts

- Empirically estimated species count always undercounts true species count
- 3 mathematical models to approximate true count
- 99% of species missed by standard lab-culture sequencing techniques (citation missing)

Conclusions

- Primitive phylogenetic event inference throughout these problems
- But they're there because they have to be! Evolutionary events guide these other areas of research
- Opportunities for phylogenetics
  - Change input of problem to short read fragments, or perhaps probabilistic sequences
  - Single reference genome replaced by population set of individual genomes motivates complex whole genome MSA

Conclusions

- Opportunities for whole-genome sequence assembly and comparison for medicine
  - Already doing detection of complex genomic events during assembly
    - Repeats
    - Transversions
    - Large insertions/deletions
- Opportunities for metagenomics
  - State of the art for alignment, tree building there is from decades ago
  - BLAST isn't perfect